Features of orality in the language of fiction: A corpus-based investigation

Jucker, Andreas H

Abstract: This paper explores the pervasiveness of features of orality in the language of performed fiction. Features of orality are typical of spontaneous spoken conversations where they are the result of the ongoing planning process and the interaction between the interlocutors, but they also occur in the context of performed fiction (movies and plays) and in narrative fiction (e.g. novels). In these contexts, they are not the result of the spontaneous planning process but are generally produced to imitate such processes. In this paper, I explore a small range of such features (contractions, interjections, discourse markers, response forms and hesitators) in four corpora of performed fiction that have recently become available (Corpus of American Soap Operas, TV Corpus, Movies Corpus and Sydney Corpus of Television Dialogue) and compare their frequency patterns with spontaneous face-to-face conversations in the Santa Barbara Corpus of Spoken American English and with narrative fiction and academic writing in the Corpus of Contemporary American English (COCA). The results confirm that the selected features of orality are used regularly in performed fiction but less frequently than in spontaneous face-to-face interactions while they are rare in narrative fiction and almost entirely absent in academic writing. The results also show that the status of the transcriptions contained in these corpora needs to be assessed very carefully if they are to be used for a study of pragmatic features.

DOI: https://doi.org/10.1177/09639470211047751

Article

Features of orality in the language of fiction: A corpus-based investigation

Language and Literature
2021, Vol. 0(0) 1–20

DOI: 10.1177/09639470211047751
journals.sagepub.com/home/lal

Andreas H. Jucker
English Department, University of Zurich, Zurich, Switzerland

Abstract

This paper explores the pervasiveness of features of orality in the language of performed fiction. Features of orality are typical of spontaneous spoken conversations where they are the result of the ongoing planning process and the interaction between the interlocutors, but they also occur in the context of performed fiction (movies and plays) and in narrative fiction (e.g. novels). In these contexts, they are not the result of the spontaneous planning process but are generally produced to imitate such processes. In this paper, I explore a small range of such features (contractions, interjections, discourse markers, response forms and hesitators) in four corpora of performed fiction that have recently become available (*Corpus of American Soap Operas*, *TV Corpus, Movies Corpus* and *Sydney Corpus of Television Dialogue*) and compare their frequency patterns with spontaneous face-to-face conversations in the *Santa Barbara Corpus of Spoken American English* and with narrative fiction and academic writing in the *Corpus of Contemporary American English* (COCA). The results confirm that the selected features of orality are used regularly in performed fiction but less frequently than in spontaneous face-to-face interactions while they are rare in narrative fiction and almost entirely absent in academic writing. The results also show that the status of the transcriptions contained in these corpora needs to be assessed very carefully if they are to be used for a study of pragmatic features.

Keywords

Corpus pragmatics, features of orality, movies, performed fiction, television series

Corresponding author:
Andreas H. Jucker, English Department, University of Zurich, Plattenstrasse 47, Zurich 8032, Switzerland.
Email: ahjucker@es.uzh.ch

1. Introduction

Fictional language is full of conversational exchanges. In plays, movies, television series and so on, conversations between characters are usually the very essence of what is going on, and even novels regularly contain large stretches of spoken interactions between characters. In some movies or television series, the conversation may appear to be very life-like, spontaneous and modelled on everyday interactions. In other cases, the interactions appear to be more stylised, for instance in a play in which characters talk in rhymed verses. But generally speaking, fictional language is scripted language, except for improvisation theatre, in which the actors themselves spontaneously create their fictional interactions as they go along (see Landert 2021). If fictional language sounds natural and life-like, it has been designed that way by the authors, scriptwriters and characters. But even in their most life-like instantiations, fictional interactions generally differ considerably from actual everyday interactions (see Bublitz 2017).

Everyday interactions contain elements that bear witness to the planning process that is going on while the utterances are produced. Often there are hesitations in the form of pauses or so-called filler words like *uh* and *um*. There are corrections, constructions that are not finished or re-started. I will have more to say on the nature and inventory of such elements later on. At this point I merely want to highlight the obvious and well-reported fact that there are significant differences between spoken and written language (Biber 1988; 2011; Cornbleet and Carter 2001; Hughes 1996; Ikeo 2019; Oesterreicher 1997). There appear to be a range of elements that are typical of spoken language, and in a preliminary and rather fuzzy manner I refer to these features as features of orality. They serve different functions, but generally and in the most typical cases they are somehow linked to the online production process, that is, the fact that speakers plan and produce their utterances more or less at the same time (see Crystal 2019: 309).

While speakers speak, they are planning what to say next and how to put that into words. In addition, they react in real time to what is happening around them, perhaps in the form of a sudden change in the context – a knock on the door, a mobile phone ringing, an object falling to the floor, or in the form of their interlocutors' reactions with a quizzical look, a gesture of disapproval, an encouraging back channel or an angry interruption. This is what I want to call the online nature of language production. Transcriptions of spoken language in many ways bear witness to this.

In the case of language in performed fiction such as movies or theatre plays (again with the exception of improvisation theatre), the online planning process for the actors is different. They do not have to think about what to say next since they have memorised their lines, and they already know what the reactions of their interlocutors will be. If their utterances appear spontaneous and unscripted, it is because they are performed in that way. In narrative fiction, which here is understood as written narrative fiction as for instance in a novel, the need for online planning disappears entirely. Written sentences are planned ahead of their transmission to the audience, and, therefore, written language might be assumed to be free of features of orality. But they also make their appearance in various forms of written language, where they are not caused by the online planning process, but they are produced in order to imitate this planning process in the speech of the characters.

This paper, therefore, wants to trace a range of features of orality in different fictional contexts and compare their frequency and use in these contexts with their occurrence in spontaneous, non-fictional interactions. In previous research, similar comparisons were usually carried out on the basis of one single genre of fictional language, for example, television dialogues (see in particular Bednarek 2018: chapter 7; and Bublitz 2017 for an overview). However, recently a range of new corpora has become available with material from television series and movies (*Corpus of American Soap Operas* (SOAP), the *TV Corpus*, the *Movie Corpus* and the *Sydney Corpus of Television Dialogue* (SydTV); for details see Section 3 below). Mark Davies, the creator of the first three of these corpora, describes them on their website as 'a great resource to look at very informal language'. In this paper, I want to put this claim to the test and check its veracity with respect to a number of selected features of orality. For this purpose, I compare the frequencies of these elements in the four TV/movie corpora with their frequency in three additional corpora (or rather corpus sections). On the one hand, I use the *Santa Barbara Corpus of Spoken American English* (SBC), a corpus of spontaneous, everyday interaction in order to see how fictional interactions in a variety of contexts compares to non-fictional interaction. On the other hand, I use two corpus sections of the *Corpus of Contemporary American English* (COCA), the Fiction section and – to get as far away from spoken language as possible – the section containing academic writing. With these corpus sections, I can compare the frequency of the selected features in performed fiction with their frequency in written fiction.

For the sake of comparability, I have chosen corpora or sections of corpora that are labelled as North American English. In the relevant corpora, this combines US English and Canadian English. This does not come with any claims that there are no differences between American English and Canadian English, but due to the structure and composition of these corpora it is not possible to separate the two varieties. And, of course, we have to be aware that movies that are labelled as North American English may also contain the speech of characters who speak other varieties of English.

In Section 2, I provide a brief overview of relevant literature as a background to a more fine-grained model of the language of fiction that distinguishes between the different forms on the basis of their comparability to spontaneous interactions in terms of language planning and production. In Section 3, I introduce the range of elements that I have used for this investigation, that is, contractions and different types of inserts (Biber et al., 1999: 1082–1099). Section 4 introduces the seven corpora (or corpus sections) and the methods of analysis. Section 5 presents the relevant frequency figures across the seven corpora, and Section 6, finally, discusses the significance of the results and concludes with some thoughts on further research opportunities in this area.

2. Orality in fiction

The orality of the language of fiction has been investigated in many different ways, both by linguists and literary theorists. Short (1996: 173–186), for instance, discusses in detail some of the essential differences and similarities between dramatic dialogue and conversations, most of which have to do with the fact that dramatic dialogue is 'written to be spoken' while casual conversation is generally unprepared and unrehearsed. As a result,

Short argues, normal non-fluency features, such as voiced fillers, silent pauses, mispronunciations, unnecessary repetitions or abandoned grammatical structures, do not occur in drama dialogue. Among the similarities between dramatic conversation and what he calls 'real conversation', he lists the turn-taking conventions and the fact that generally people do not talk at the same time, apart from short overlaps. He also mentions the speech act values of utterances that are roughly the same in fictional and non-fictional contexts and the fact that people often speak indirectly and say something while they mean something else. As a case in point, he compares short extracts taken from two plays, one by Harold Pinter and the other by George Bernard Shaw, and speculates on the extent to which they sound like spontaneous conversations.

Thomas (1997, 2002) uses the descriptive framework of conversation analysis to get a better understanding of specific fictional texts by Evelyn Waugh. In one case, she investigates Waugh's novel *Vile Bodies* and dissects the depicted telephone conversations between the two protagonists with conversation analytical tools (Thomas 1997), and in the other, she analyses the intricacies of multiparty talk in Waugh's *Black Mischief* (Thomas 2002). These studies also reveal a range of important similarities and differences between fictional dialogues and spontaneous interactions. In the case of the multiparty talk in Black Mischief, for instance, she points out how difficult it is to keep track of who is speaking in a sequence of utterances that are not specifically assigned to individual speakers by the narrator. And at the same time the interaction reveals the power play that is going on between the characters who take part in the conversation. In an important handbook article, Lambrou (2014) extends the scope of linguistic approaches that are brought to bear on an analysis of fictional discourse. In addition to conversation analytical tools, she also discusses the usefulness of Grice's (1975) cooperative principle in the interpretation of fictional dialogues, as for instance in the case of unexpected replies and other floutings of the conversational maxims.

Other researchers have set out to substantiate the differences between fictional and non-fictional interactions not just on a qualitative basis but also with wider ranging quantitative claims based on corpus linguistic methods. These approaches are particularly important as a background for my own research to be presented below. A recent example is Ikeo (2019), who uses lexical categories (personal pronouns, proper nouns, determiners, etc.) to investigate the relationship between written fiction and spoken discourse. She compares two small corpora of recent novels; one containing present-tense narratives and the other containing past-tense narratives, and she finds that for all categories the present-tense narratives show frequencies that are closer to the frequencies of these elements in spoken language as reported by Biber et al. (1999). Personal pronouns, for instances, comprise 10.48% of all the words in the present-tense corpus while the past-tense corpus shows a frequency of only 8.45%. This accords with Biber et al.'s (1999: 235) findings that pronouns are used more frequently in conversations than in written language. Proper nouns, on the other hand, are more frequent in the past-tense corpus (4.38%) than in the present-day corpus (3.75%). She concludes that 'present-tense narrative seems to be closer to spoken discourse than past-tense narrative in the following ways, having (1) more finite verbs and verb phrases, (2) more pronouns, (3) fewer proper nouns and adjectives and (4) more present progressives' (Ikeo 2019: 300), and she interprets these findings as an

indication of the often-observed recent tendency towards colloquialisation (see Leech et al., 2009).

Bednarek (2018) is a more detailed study with a larger scope. She looks at words and n-grams and compares their relative frequencies in the SydTV with that in five reference corpora of different types of spoken English, both American and British, and she uses Sketch Engine's functionality to establish the similarity between each of these six corpora. This is based on the hypothesis that '[i]f TV dialogue is a partial or selective imitation of unscripted spoken interaction, SydTV-Std should be more similar to corpora that contain such interactions (i.e. non-specialised spoken corpora) than to corpora that contain written language' (Bednarek 2018: 122). It turns out that the *Longman Spoken American Corpus* with its five million words of American English conversations (Biber et al., 1999) is most similar to the SydTV in terms of Sketch Engine's *Comparing Corpora* function. Bednarek (2018: 123) takes this as evidence that 'TV dialogue successfully imitates unscripted conversation'. The shared n-grams that appear to be responsible for the close similarity between these two corpora include a long list of categories, for instance references to time and place (e.g. *right now*, *in there*), discourse markers (e.g. *I mean*, *you know*, *what I'm saying*), routine formulae (e.g. *thanks for*, *see you later*), interjections (e.g. *oh god*, *oh man*), n-grams that include a first-person or a second-person pronoun (e.g. *I don't even know*, *let me tell you*), and so on. She concludes that 'these n-grams help to construct TV dialogue as if it was unscripted talk' (Bednarek, 2018: 125) and refers to Heyd (2010) notion of 'staged orality', that is, the use of elements that are typically perceived to be 'oral, spontaneous, or intimate' (2010: 34).

In an earlier study, Quaglio (2009) compares the language of the television series *Friends* with conversational data on the basis of Biber's (1988) multidimensional analysis (see also Biber et al., 2002). This approach is based on a large number of linguistic features that previous studies have shown to be associated with different functions of language. The selection of features includes, for instance, first- and second-person pronouns, contractions and demonstrative pronouns, which have been associated with conversations. Moreover, it includes the use of passives and nominalisations, which have been associated with academic writing (Quaglio, 2009: 58). These features are then automatically identified in a large range of registers in order to get fine-grained and multidimensional profiles of all these registers. In his investigation, Quaglio compares a corpus containing approximately 590,000 words of transcripts of nine seasons of the series *Friends* with the results obtained by Biber (1988) for face-to-face conversations. He finds that the statistics derived through the most important dimension of Biber's investigation, dimension 1 of involved versus informational production, produces almost identical results (Quaglio 2009: 65), but interestingly it turns out that the variability in Biber's face-to-face conversational data is larger than the variability in his *Friends* data. He concludes:

> Therefore, the score obtained by *Friends* on D1 indicates that the language of the television show is similar to face-to-face conversation from a grammatical point of view. In other words, the co-occurrence patterns of linguistic features in *Friends* are similar to those typifying face-to-face conversation. (Quaglio, 2009: 68)

All these approaches, with different levels of sophistication and comprehensiveness, use lexical items, lexical categories and their co-occurrence patterns as a starting point for their investigation of the difference between the language of fiction and spontaneous spoken interaction. Other approaches try to identify elements that are specifically associated to spoken language through their functions that have to do with the online planning and production of speech, the so-called features of orality. Bublitz (2017: 243), for instance, proposes five large-scale categories to account for the diversity of different features of orality. The first category comprises features of metacommunication under which he subsumes those elements of spontaneous face-to-face conversations that refer to the communication process itself through clarifying, paraphrasing and repairing what has been said. Pertinent examples are discourse markers, such as *well*, *oh*, *I think*, *you know*, or *like*, hesitators, disfluency markers or planners, such as *er* and *erm* (or *uh* and *um* in American English spelling), and repetitions. Bublitz' second category comprises features of turn management, in particular all those elements that are used as turn-taking signals or otherwise contribute to the cooperation of speaker and hearer in the joint construction of discourse, for example, back-channel signals or listener responses. The third category consists of features of topic management and includes, for instance, digression indicators. The first three categories are primarily concerned with the planning and repairing of discourse and its organisation. The fourth category combines features of involvement, that is, features that are related to the fact that speakers get involved with their interlocutors where writers are more detached from their readers. This category cuts across most oral features typical of the three preceding categories and stands in opposition to features of detachment that are more typical of written language. And the fifth category, finally, includes features of formal reduction, which can be phonetic contraction or incomplete syntax.

In this paper, I want to combine the two approaches by using a carefully selected range of features of orality (along the lines of Bublitz 2017) as a basis for a comparison of corpora that contain fictional data with reference corpora containing spontaneous spoken interaction on the one hand and written language on the other (along the lines of Bednarek, 2018; Ikeo, 2019 or Quaglio, 2009). Features of orality, in a very general sense, therefore, are those features that are associated with the online and spontaneous production of utterances. They are related to the planning process and to the interaction between speaker and addressee and the way in which speaker and addressee take turns at speaking. In order to understand the distribution of these features, it is necessary to distinguish carefully between different levels of planning and production. Figure 1, taken from Locher and Jucker (2021), provides an overview. It distinguishes between performed fiction and written fiction. Performed fiction involves actors who perform specific characters and their interactions. Written fiction consists of written texts, in which the audience can find written versions of the interactions between the fictitious characters. In the case of play texts or movie scripts, the dialogues between the characters generally take up most of the written text. In the case of narrative texts, there may be very little dialogue between the characters or there may be almost as much as in a play text. But it is probably fair to say that narrative fiction in the form of novels or

Type of text	**Spontaneous interaction**	**Performed fiction**			**Written fiction**	
Subtype		Improv theatre	Plays	Movies, TV series	Play texts, movie scripts	Dialogues in narrative texts
Planning	Online		Offline			
Production	Online			Frozen	None	
Type of orality	Genuine		Staged		Represented	
Reasons for orality features	Orality features largely consequence of online planning and production		Orality features partly consequence of online production, partly used for narrative purposes		Orality features used for narrative purposes	

Figure 1. Orality in spontaneous interaction, performed fiction and written fiction (Locher and Jucker 2021: 136).

short stories regularly includes a reasonable amount of interaction between the depicted characters.

As pointed out above, fictional language is typically planned before its production. It is a case of offline planning, except for improvisation theatre in which the actors plan what to say while they are performing. They may use cues or keywords provided by the audience to improvise some scenes, and thus in terms of planning, their performance is more like spontaneous interaction than like the performance of a play that is performed on the basis of a written text that has been memorised by the actors. In movies and TV series, the actors typically also produce interactions that have been memorised and rehearsed beforehand, even though this may include varying amounts and degrees of improvisation. However, in contrast to improvisation theatre and plays, the production is not really online. It is not produced while the audience watches it, but it is a case of what Locher and Jucker (2021) call frozen production.

On this basis, I can now specify more precisely what orality really means in these cases. In spontaneous interactions and in improvisation theatre, orality is the result of online planning and production. Hesitations, corrections, interruptions and so on may be caused directly by the problems in the planning process, but of course even in these situations, some of these features may be produced for effect. For example, an interactant in a non-fictional encounter may strategically use backchannels to signal turn-taking intent or hesitation markers to foreshadow a dispreferred response. An improvisation artist might be taken aback by what a co-actor said and respond with surprise tokens. In both cases the parties do not know exactly what the next contributions will be like although there are educated guesses due to the frame within which the interaction takes place. In the case of enacted drama, an artist knows what comes next.

3. Contractions and inserts as features of orality

In order to investigate the level of orality in different contexts, I selected a number of features that are strongly associated with spoken language. For practical reasons, they had to be characterised by some additional conditions to be included in the selection. First, they had to be relatively frequent, at least in the reference corpus of spoken language. And second, they had to be electronically retrievable, which means that the search strings had to be reasonably simple in order to be compatible with the search interfaces of the different corpora. The elements that correspond most closely to this description are contractions and a range of inserts (Biber et al., 1999: 1082–1099).

In English, there are two types of contractions; verb contractions (as in 'they're playing chess') and contractions of *not* (as in 'they wouldn't play chess'). Biber et al. (1999: 1128–1132) provide detailed statistics on the occurrence of these contractions in four different registers; conversation, fiction, news and academic writing. The rates vary for different verbs and different syntactic positions, but the main difference in the frequency of contractions is caused by the register in which they occur. In conversations, they are very frequent. In this register, the verbs *be* or *will* are contracted in about 75% of all cases, in contrast to less than 2% in academic writing. In fiction writing, the percentage of contractions for these two verbs reaches some 45% to 50%, and in news writing some 5% to 10%. Negative contractions follow the same pattern, but they are somewhat more frequent than verb contractions in all four registers. After the verb *do* or a modal verb, *not* is contracted in nearly 100% of all cases in conversations. In fiction writing, the frequency of *not*-contraction after *do* or a modal verb reaches 75% and 65%; in news 60% and 40%, and in academic writing only about 5% in both syntactic positions. Biber et al. (1999: 1129) explain these patterns with the level of admixture of spoken style in the form of direct reporting of spoken discourse, which is particularly high in fiction writing. It also occurs in news, but it is rare in academic writing. In the statistics presented below, I did not compare the contractions with their uncontracted variants, instead I calculated the frequency of contractions per million words for each corpus (or corpus section). For practical purposes only a selection of very common contractions was included, as not all the corpora under investigation allowed a direct search for all contractions.

Inserts, according to Biber et al. (1999: 1082, highlight removed) are 'stand-alone words which are characterized in general by their inability to enter into syntactic relations with other structures'. Prototypically, they are peripheral elements both in terms of their grammatical behaviour and in terms of their lexical shape. Interjections (like *oh*), discourse markers (like *well*, *you know*, or *I mean*), response forms (like *yes* or *yeah*) and hesitators (like *uh*/*er* and *um*/*erm*) are the most common inserts in American English and British English conversation (Biber et al., 1999: 1096), but the different categories are not clearly separated from each other. Specific functions often overlap. Inserts tend to be morphologically simple, and typically, they do not have homonyms in other word classes (discourse markers are an exception), and they are defined through their pragmatic function rather than their semantic meaning.

Interjections have 'an exclamatory function, expressive of the speaker's emotions' (Biber et al., 1999: 1083). The two most frequent ones according to Biber et al. (1999: 1083) are *oh* and *ah* (see also Aijmer, 1987; Norrick, 2011; 2015; Stange, 2016). The

function of interjection overlaps to a large extent with that of discourse markers, and, in fact, many researchers treat *oh* as a discourse marker (e.g. Schiffrin, 1987) or a (discourse) particle (e.g. Aijmer, 1987, 2002; Heritage, 1998, 2002). For the purpose of this article, I adopt Biber et al.'s classification, and I included the interjections *oh* and *ah*.

Discourse markers, according to Biber et al. (1999: 1085) are 'inserts which tend to occur at the beginning of a turn or utterance, and to combine two roles: (a) to signal a transition in the evolving progress of the conversation, and (b) to signal an interactive relationship between speaker, hearer, and message.' In contrast to interjections, discourse markers regularly have homonyms in other word classes with non-discourse marker functions. The most common ones are *well*, *you know* and *I mean* (Biber et al., 1999: 1096), and it is these that I included in this investigation (see also Furkó, 2020).

Response forms, response elicitors and back channels are elements that are particularly strongly associated with spoken interaction. They help to negotiate the turn-taking between the interactants. Response forms are 'inserts used as brief and routinized responses to a previous remark by a different speaker' (Biber et al., 1999: 1089). They include elements like *yes*, *yeah*, *okay*/*ok*, *no* or *uh huh*, which are generally used to respond to questions or requests, and so-called backchannels like *mhm*, which are used to respond to assertions. In this investigation, I included the response forms *yes*, *yeah*, *mhm*, *uh huh*, *no*, *ok* and *okay*.

Hesitators are defined by Biber et al. (1999: 1092) as 'pause fillers, whose main function is to enable the speaker to hesitate, i.e. to pause in the middle of a message, while signalling the wish to continue speaking'. In American English, they generally take the written form of *uh* or *um*, in British English they are usually spelled *er* and *erm*. Several additional spelling variants exist for them. The transcribers of the SydTV regularly used the spelling *umm*. Many researchers prefer the term planner for these elements because of its more positive connotations (e.g. Jucker, 2015a; 2015b; Staley and Jucker, 2021; Tottie 2011; 2014). In the context of this paper, however, I retain Biber et al. (1999) terminology and call them hesitators.

Table 1 gives an overview of all the elements that were retrieved for the statistical comparison. In addition to the contractions, this selection of items mostly follows the list of what Biber et al. (1999: 1096) describe as the most common inserts for American and British English conversations. As pointed out the categories are not clear-cut, and functional overlaps occur. The elements are listed under their most common category, and

Table 1. Selected features of orality.

Contractions	
Verb contractions	*that's, it's, I'm, you're*
Negative contractions	*don't, doesn't, wasn't, weren't*
Inserts	
Primarily interjections	*oh, ah*
Primarily discourse markers	*well, you know, I mean*
Primarily response forms	*yes, yeah, okay/ok, no, uh huh, mhm*
Hesitators	*uh, um, uhm, u=m, umm* (in SydTV for *um*)

in the quantitative corpus analysis I made no attempt to distinguish between different uses of the same element. *Okay* for instance can be used both as a response form, when it is a reaction to what an interlocutor has just said, and as a response elicitor, when it is used to elicit an agreement (or disagreement) from the addressee. All searches were checked and – if necessary – manually adjusted to exclude false hits (see below, Section 3).

The features of orality in Table 1 have in common that they have a very clear interactive component and that they are part of the online processing of spoken language. The list is far from complete, but it contains elements that are reasonably easy to retrieve via different corpus interfaces, and as a set they serve as diagnostics for my research questions. The investigation is based on the assumption that different, or indeed more comprehensive, sets would provide roughly comparable results, but to what extent this assumption is correct must remain an open question.

4. Data and method

The current investigation focuses on the language of performed fiction. For this purpose, four corpora were used that all record different types of spoken interactions depicted in movies and in television series. In order to put these representations into perspective, I compare them, on the one hand, to spontaneous unscripted interaction and on the other to written language in the form of fiction writing and in the form of academic writing. All corpora contain US data or in some cases North American data. Some corpora also incorporate material from other national varieties of English, but in these cases, I restricted the searches to the North American, that is, US and Canadian, data in order to increase the overall coherence of the data. A further restriction to US data only was not possible because the relevant corpora do not distinguish between US and Canadian English. And again, it is important to remember that movies often depict characters who speak other varieties of English.

The television and movie data consist of the SOAP, the *TV Corpus*, the *Movie Corpus* and the SydTV. The SOAP contains 22,000 transcripts from American soap operas from the early 2000s amounting to 100 million words. The *TV Corpus* contains 325 million words from 75,000 TV episodes from the 1950s to the current time. The subpart labelled US/CA used for this investigation consists of 266 million words. The *Movie Corpus* contains 200 million words from 25,000 movies from the 1930s to the current time. In this case, the relevant subpart of US/CA transcripts amounts to 153 million words. It is possible, of course, that the historical dimension of the *TV Corpus* and the *Movie Corpus* influenced the results to some extent, but for practical purposes this dimension was not explored further together with several other dimensions that might also have played a role in the frequency of features of orality, such as genre differences, gender differences, differences of depicted social classes and so on.

According to Mark Davies (personal communication), the TV and Movies corpora are based on the subtitles retrieved from the website OpenSubtitles.org. He assumes that the subtitles are a closer representation of what is actually said than the screenplays. The data for SOAP have been taken from http://tvmegasite.net/day/transcripts.shtml and seems to be based on volunteer transcriptions. For pragmatic analyses, such uncertainties are problematic, and I would certainly want to question the accuracy of the subtitles as

transcriptions since they are subjected to serious constraints of space and time (Frumuselu et al., 2015; Guillot, 2012; 2017). They need to be shown for a sufficient amount of time in order to be readable and they have to fit into the character count that is available on the screen. It is reasonable to assume that at least some oral features, for example, hesitators, may be left out in subtitles for space reasons. As pointed out above, the Mark Davies website advertises these corpora 'as a great resource to look at very informal language'. However, I want to stress that the informality of the interactions in these corpora is, of course, a scripted informality.

The SydTV[1] is more accurate and linguistically more reliable than the corpora on the Mark Davies website, but with about a third of a million words, it is also much smaller than these. It was compiled by Monika Bednarek together with her team who either transcribed all the episodes contained in the corpus from scratch or corrected existing transcriptions on the basis of the actual movie (see Bednarek 2018: chapter 5 and appendix for details).

To these corpora of performed fiction I have added two reference corpora. *The Santa Barbara Corpus of Spoken American English* (SBC) has been used as a sample of spoken language. It is relatively small and consists of some 250,000 words of transcriptions of interactions, such as conversations between family and friends, task-related interactions in a small shop, a veterinarian office or an air traffic control tower, but also lectures and sermons. The interactions were recorded in the years 2000 to 2005 (Brown 2015; Du Bois et al. 2000-2005).

The second point of reference is provided by corpus sections containing written language. For this I used the COCA, and in particular the Fiction section and the Academic writing section. At the beginning of this investigation in late 2019 and early 2020, the corpus consisted of 560 million words. In March 2020, COCA was considerably extended. Three new genres were added (Blogs, Web and samples from the TV Corpus and the Movies Corpus), and all sections were updated to also include material for the years 2018 and 2019. This increased the total corpus size from 560 million words to one billion words. The statistics for the Fiction section are based on the 2019 version which at the time consisted of 112 million words from the 1990s up to 2017. For the Academic writing, which was investigated somewhat later, the statistics are based on the 2020 version, which now consists of 121 million words from 1990 to 2019. The Fiction section contains short stories and first chapters of first edition books, children's magazines and popular magazines. However, it also contains a certain amount of performed fiction in the form of plays from literary magazines and movie scripts. Thus, the distinction between the performed fiction of the main corpora and the written fiction in the COCA Fiction section is not as clear-cut as I might have wished. The material for the section with academic writing is drawn from nearly one-hundred different academic journals across a wide range of academic specialisations. This genre was added to provide a baseline of texts in which no features of orality would be expected. Their non-occurrence in this genre, therefore, provides some additional evidence for their oral nature. Table 2 provides an overview of the corpora used for this study. The information on the individual corpora was taken from Brown (2015: 51) for SBC; Bednarek (2018: 82) for SydTV; and from the Mark Davies website for all others.

Table 2. Composition of data.

Spontaneous interaction			
Santa Barbara Corpus of Spoken American English (SBC)	2000–2005	249,000 words	Spoken conversations
Performed fiction			
Corpus of American Soap Operas (SOAP)	2001–2012	100 mio words	TV shows
The TV Corpus (US/CA)	1950–2018	325 mio words	TV shows
The Movie Corpus (US/CA)	1930–2018	200 mio words	Movies
Sydney TV Corpus (US TV series)	2000–2012	334,000 words	TV shows
Written fiction			
Corpus of Contemporary American English (COCA) (Fiction)	1990–2017	112 mio words	Narrative fiction
Non-fiction (academic)			
Corpus of Contemporary American English (COCA) (Academic)	1990–2019	121 mio words	Academic written

In order to establish the frequencies of all the features of orality listed in Table 1 above, they were retrieved individually in each of the corpora listed in Table 2. In most cases, the searches also retrieved hits that were not orality related. A search for the discourse marker *well*, for instance, also retrieved passages where it was used as a regular adverb or a noun. A search for the interjection *oh* also retrieved passages where it was used as an alternative expression for zero, and so on. On the basis of the retrieved hits, it was therefore necessary to establish a reasonably accurate estimate of real hits versus false hits. To achieve this, the following procedure was used.

In a first step, the raw figures were established. For the corpora contained in the English Corpora website and for the SydTV the respective retrieval software offered by these websites was used. In the case of the *Santa Barbara Corpus*, the hits were retrieved with LancsBox 4.5. As pointed out above, the searches in the TV Corpus and the Movie Corpus were restricted to the US/CA section, that is, to data from the US and Canada. In the COCA, the searches were carried out in the Fiction section and in the section containing academic writing.

In a second step, random samples of each search were used to establish the level of noise (i.e. the frequency of false hits) for each search string. For each item, a trained coder manually inspected two independent sets of 100 random hits and established the percentage of real hits versus false hits for each of the two sets. If the difference between percentages for the two sets was less than 15 percentage points, the average of the two values was accepted as reasonably accurate value for the given search string. Otherwise, a third set of 100 random hits was inspected. However, this turned out to be necessary only in one case (for *you know* in SydTV). These figures were then used to calculate reasonably accurate frequencies of each feature of orality in each of the investigated corpora. This procedure does not lay claim to absolute precision, but the fact that all searches were carried out by the same coder, and that for every single search two independent samples were coded assures a sufficient level of reasonable accuracy.

5. Frequency distribution

The four main corpora of this investigation (SOAP, TV Corpus, Movie Corpus and SydTV) all contain records of scripted and performed interactions. They are based on fictional language produced for an audience (see Locher and Jucker, 2021; Locher, 2017). But as pointed out above, the four corpora differ in the ways in which the spoken interactions were transformed into written data.

Figure 2 provides a breakdown of the frequency figures across seven (sections of) corpora. It contrasts the total frequency of the selection of inserts listed in Table 1 above with the frequency of contractions. The figure clearly shows that contractions behave differently from inserts.

The *Santa Barbara Corpus* shows the highest incidence of inserts with 41,681 instances per million words. SOAP follows on second place while the remaining three movie corpora show a somewhat smaller frequency. In the fiction part of COCA, on the other hand, the frequency of inserts is very small, with 2175 instances per million words. In academic writing, they are virtually non-existent with only 107 instances per million words. This is not entirely unexpected. In the spontaneous conversations contained in the SBC, the high frequency of such features can easily be predicted. It is interesting to see, however, that in the fictional spontaneity of performed fiction, these features are less prevalent. It appears that the fictional interactions rely to a lesser extent on the interactivity and orality of inserts. In narrative fiction, the frequency of inserts is much lower. In this

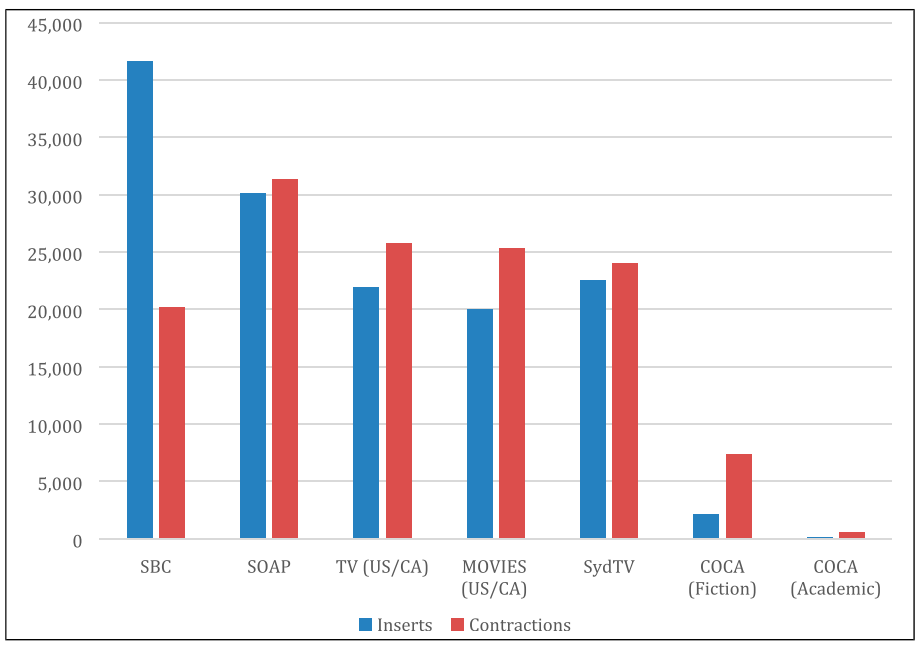

Figure 2. Frequency of common inserts and contractions in six different corpora (per million words; manually adjusted to exclude false hits).

particular respect, narrative fiction is much closer to academic writing than to the performed fiction in movies, and it is reasonable to assume that the frequency of inserts is closely related to the frequency of direct reported speech.

Contractions, on the other hand, do not have their highest frequency in the *Santa Barbara Corpus*. The four movie corpora all show higher rates of contractions. They range between 24,069 per million words in the *Sydney TV Corpus* and 31,735 per million words in SOAP, in contrast to only 20,145 instances per million words in the *Santa Barbara Corpus*. It is again according to expectations that the two written samples manifest much lower frequencies, but in this case, there is a much clearer difference between COCA Fiction with 7332 instances per million words and COCA Academic with only 551 per million words. Presumably, contractions are not as closely connected to reported speech as inserts.

A comparison of the frequency of inserts and contractions is of limited value only because the frequencies depend on the selection of elements that were included in each category. On the basis of Biber et al.'s (1999) figure and a large range of trial searches, the selection appears to include all of the most common inserts, but the list of elements is not exhaustive, and the inclusion of additional inserts might modify the comparison as it presents itself in Figure 2. However, it is striking that the frequencies for inserts and for contractions are roughly comparable for the four movie corpora while they differ more significantly both in the *Santa Barbara Corpus* and the two corpus sections containing written language. In the *Santa Barbara Corpus*, there are more than twice as many inserts as contractions (41,681 vs 20,145 per million words), while in the written corpus sections the relation is the other way round. Here, the contractions outnumber the inserts by 3.4 to 1 and by 5.1 to 1 (7332 vs 2175 in COCA Fiction, and 551 vs 107 in COCA Academic).

Let us now have a closer look at the different types of inserts in the seven different corpora (sections). Figure 3 provides an overview of their frequencies. When interpreting Figure 3, it must be remembered that each category is represented by no more than a selection of its most common members. None of these categories is exhaustive, and, as pointed out above, there is even some functional overlap between the categories. But the difference across the corpora or corpora sections is interesting. Each category shows its highest frequency in the *Santa Barbara Corpus* and very low frequencies in the two written COCA sections. The four movie corpora show frequencies that lie between these two extremes.

Apart from these very general observations, there are a few more specific ones. First, the frequency of response forms in SOAP is almost as high as that in the *Santa Barbara Corpus* and much higher than the other three movie corpora. It is difficult to say whether this is due to a higher level of interactivity of the material included in this corpus or whether this has anything to do with the way in which these interactions were turned into written texts. In the second case, this appears to be somewhat clearer. The *Sydney Corpus TV Dialogues* manifests a much higher level of hesitators than the other three corpora. In this case, the different transcription principles might be the main reason for the difference. As pointed out in Section 3 above, SydTV consists of careful transcriptions produced or adapted by trained linguists, while the other three corpora are based on subtitles. It is to be expected that trained linguists would be careful to include hesitators in their transcriptions, while the producers of subtitles might be much more likely to consider them as irrelevant

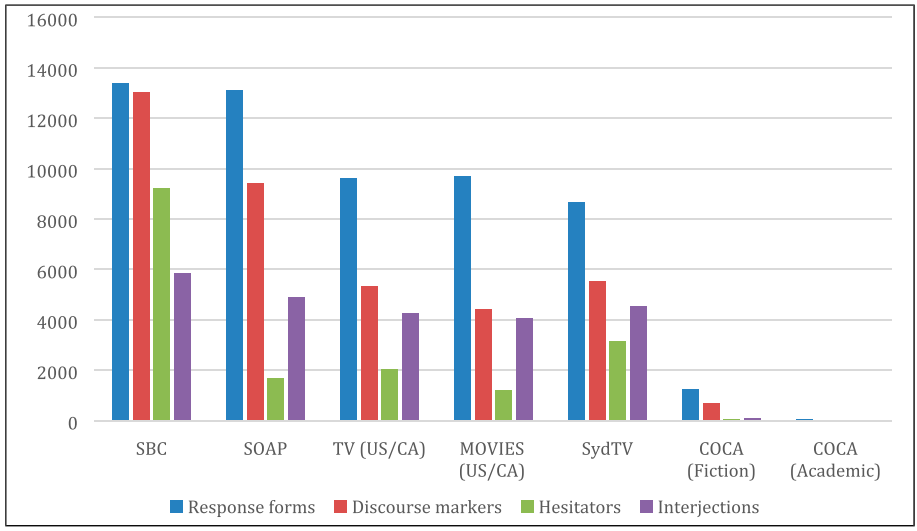

Figure 3. Frequency of four categories of common inserts in six different corpora (per million words; manually adjusted to exclude false hits).

or even distracting and therefore exclude them from the subtitles, which at the same time reduces the character count for each subtitle. In this light, it is interesting to note that the same cannot be said for response forms, discourse markers and interjections. It might have been expected that in subtitles, where space is at a premium, they would tend to be omitted while linguistically trained transcribers would be careful to include them. But it appears that only hesitators are regularly omitted from subtitles whereas response forms, discourse markers and interjections are sufficiently important to be retained in the subtitles.

6. Conclusion

The investigation presented above has focused on a small range of elements that are closely associated with spoken language: contractions and a few common inserts. I have used them as diagnostics to compare the language of performed fiction with spontaneous spoken interactions on the one hand and with written language in the form of narrative fiction and in the form of academic writing on the other. The results of the investigation are interesting both in terms of the different types of language under analysis and in terms of the corpora under investigation.

At first sight, the frequency distribution of these features of orality very much follows the expected pattern. They are most frequent in the spontaneous face-to-face interaction contained in the *Santa Barbara Corpus*. They are almost entirely absent in academic writing, and they are relatively rare, but they do occur in the written language of the COCA written fiction section. The four corpora containing movies and television series show frequencies that are closer to the spoken language than to the written language. But the results also reveal some less obvious insights. As mentioned above, Mark Davies

advertises SOAP, the *TV Corpus* and the *Movie Corpus* as sources of very informal language. The term 'informal' is, of course, somewhat fuzzy and hard to pin down exactly. But it is clear that in terms of the features of orality investigated here, the scripted language of fiction of movies and TV series differs significantly from spontaneous face-to-face conversations. It is certainly closer to spoken interactions than written language, but it cannot serve as a substitute for spontaneous conversations, nor should it be expected to. The language of fiction is subject to its own communicative constraints and needs to be investigated in its own right. In performed fiction, features of orality are – generally speaking – not part of the spontaneous online planning and production process. They are part of 'staged orality' (Heyd 2010). The actors who perform characters in movies and TV series perform orality, and it appears that the performance of orality can be achieved with a reduced frequency of features of orality. The features serve as indexical cues (in the sense of Bucholtz and Hall 2005: 594 or Locher and Jucker 2021: 99) and conjure up a sense of orality without having to mirror spontaneous orality in all its complexity, in the same way that in a historical movie a few phrases like 'anon' or 'I like her not' may be enough to conjure up Early Modern English even if they are embedded in Present-day English (Locher and Jucker 2021: 105).

Bednarek (2018: 23) argues that 'most TV dialogue artfully but selectively simulates naturalistic speech'. According to her (2018: 23), TV dialogue:

- favours comprehensibility/intelligibility (e.g. is more clearly enunciated, is less vague)
- tends towards focus, coherence, fluency (e.g. is less narrative, has different turn lengths and organisation, has fewer interactive and performance features)
- has a focus on emotionality and entertainment (e.g. is more emotional, may focus on conflict or humour, may feature exaggerated/stereotypical language use)
- permits the use of devices that foreground the 'constructedness' of the dialogue.

These features may lead to overuse of some linguistic features and underuse of some others. The investigation above indicates that features of orality appear to be underused, and it is easy to speculate that a higher incidence would reduce the comprehensibility and intelligibility of what is being said by the characters as well as the overall coherence. It would most likely also slow down the interaction and – considering the generally quick style of interaction with few breaks and pauses in present-day movies and TV series – it may even be speculated to put off some viewers who would perceive the interactions and characters as slow and boring.

However, my investigation has also revealed some interesting differences between the four corpora containing performed fiction. As pointed out above, the *Sydney Corpus TV Dialogues* is the most reliable in terms of the transcription principles. In this corpus, most of the investigated features show somewhat lower frequencies with the notable exception of hesitators. It seems plausible to assume that the higher frequency of hesitators is a direct result of the trained linguists who carried out and checked all the transcriptions of this corpus. They can be expected to have paid close attention to an element that has attracted so much research in the relevant research literature, whereas subtitles are much more likely

to ignore these elements as largely irrelevant or as a good way to save some space in the restrictive character count for every single subtitle.

The other notable exception is the corpus of soap operas. This corpus manifests notably higher frequencies both for response forms and for discourse markers. In the case of response forms, they almost reach the level attested for the spontaneous face-to-face interactions in the *Santa Barbara Corpus* while the levels of hesitators and interjections is very similar to those attested for the *TV Corpus* and the *Movie Corpus* (see Figure 3 above). This finding is more difficult to interpret, but in this case, the pattern may actually be related to the communicative profile of soap operas. The focus of SOAP on one specific genre of television series may highlight specific interactional features that are less prominent across a broader range of genres as they are contained in the remaining movie and TV corpora. Another possible explanation may lie in the larger time depth of the *TV Corpus* and the *Movie Corpus*. It is, of course, possible that the frequency of these features is a relatively recent phenomenon, which would give them a higher level in the more recent SOAP material than in the other two corpora which also incorporate older material.

Clearly, more research is needed to disentangle the differences between the four main corpora of this investigation and their relationships to spontaneous face-to-face interactions as they are recorded in the *Santa Barbara Corpus* and to the written genres contained in COCA Fiction and COCA Academic. These corpora provide a wealth of material of performed fiction and they open up exciting new research opportunities, but they should not be mistaken for spontaneous face-to-face interactions. The analysis of selected features of orality has underlined the need to investigate them in their own rights rather than as a substitute for something else. Features of orality occur regularly in these corpora, but they are not the result of the online production process as these features are in spontaneous conversations. They are part of the staged orality of performed fiction.

Acknowledgements

Some of the ideas for this paper go back to the joint work for Locher and Jucker (2021). My thanks go to Miriam Locher for many inspiring discussions of the issues reported in this paper and for very detailed and helpful feedback on a preliminary version of this paper. My thanks also go to Anja Leu for her help in coding the data and to the anonymous reviewers for their very helpful comments. The usual disclaimers apply.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. http://www.syd-tv.com

ORCID iD

Andreas H. Jucker https://orcid.org/0000-0003-3495-2213

Corpora

Santa Barbara Corpus of Spoken American English: https://www.linguistics.ucsb.edu/research/santa-barbara-corpus
English Corpora (SOAP, TV Corpus, Movie Corpus, COCA): https://www.english-corpora.org
Sydney Corpus of Television Dialogue: http://cqpw-prod.vip.sydney.edu.au/CQPweb/

References

Aijmer K (1987) *Oh* and *ah* in English conversation. In: Meijs Willem (ed) *Corpus Linguistics and Beyond. Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora*. Amsterdam: Rodopi, pp. 61–86.

Aijmer K (2002) *English Discourse Particles. Evidence from a Corpus*. Studies in Corpus Linguistics 10. Amsterdam/Philadelphia: John Benjamins.

Bednarek M (2018) *Language and Television Series. A Linguistic Approach to TV Dialogue*. Cambridge: Cambridge University Press.

Biber D (1988) *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber D (2011) Speech and writing: linguistic styles enabled by the technology of literacy. In: Andersen Gisle and Aijmer Karin (eds) *Pragmatics of Society*. (Handbooks of Pragmatics 5). Berlin/New York: De Gruyter Mouton, pp. 137–152.

Biber D, Conrad S, Reppen R, et al. (2002) Speaking and writing in the university: a multidimensional comparison. *TESOL Quarterly* 36(1): 9–48.

Biber D, Johansson S, Leech G, et al. (1999) *Longman Grammar of Spoken and Written English*. London: Longman.

Brown LR (2015) *A Corpus Study of Requests in Naturally Occurring Spoken American English. A Context Analysis Approach*. PhD Dissertation, University of Texas at Arlington.

Bublitz W (2017) Oral features in fiction. In: Locher MA and Jucker AH (eds) *Pragmatics of Fiction*. (Handbooks of Pragmatics 12). Berlin: de Gruyter, pp. 235–264.

Bucholtz M and Hall K (2005) Identity and interaction: a sociocultural linguistic approach. *Discourse Studies* 7(4–5): 585–614.

Cornbleet S and Carter R (2001) *The Language of Speech and Writing*. London and New York: Routledge.

Crystal D (2019) *The Cambridge Encyclopedia of The English Language*. Third Edition. Cambridge: Cambridge University Press.

Du Bois JW, Chafe WL, Meyer C, et al. (2000-2005) *Santa Barbara Corpus of Spoken American English*. Parts 1-4. Philadelphia: Linguistic Data Consortium.

Frumuselu AD, De Maeyer S, Donche V, et al. (2015) Television series inside the EFL classroom: bridging the gap between teaching and learning informal language through subtitles. *Linguistics and Education* 32: 107–117.

Furkó PB (2020) *Discourse markers and beyond. Descriptive and critical perspectives on discourse-pragmatic devices across genres and languages*. (Postdisciplinary Studies in Discourse). Cham: Palgrave Macmillan.

Grice PH (1975) Logic and conversation. In: Cole P and Morgan JL (eds) *Syntax and Semantics 3: Speech Acts*. New York: Academic Press, pp. 41–58.

Guillot M-N (2012) Film subtitles and the conundrum of linguistic and cultural representation. In: Hauser S and Luginbühl M (eds) *Contrastive Media Analysis. Approaches to Linguistic and Cultural Aspects of Mass Media Communication*. (Pragmatics & Beyond New Series 226). Amsterdam/Philadelphia: John Benjamins, pp. 101–122.

Guillot M-N (2017) Subtitling and dubbing in telecinematic text. In: Locher MA and Jucker AH (eds *Pragmatics of Fiction.* (Handbooks of Pragmatics 12). Berlin: de Gruyter, pp. 397–424.

Heritage J (1998) *Oh*-prefaced responses to inquiry. *Language in Society* 27: 291–334.

Heritage J (2002) *Oh*-prefaced responses to assessments: a method of modifying agreement/disagreement. In: Ford CE, Fox BA and Thompson SA. (eds) *The Language of Turn and Sequence*. Oxford: Oxford University Press, pp. 196–224.

Heyd T (2010) How you guys doin'? Staged orality and emerging plural address in the television series *Friends*. *American Speech* 85(1): 33–66.

Hughes R (1996) *English in Speech and Writing. Investigating Language and Literature*. London: Routledge.

Ikeo R (2019) 'Colloquialization' in fiction: a corpus-driven analysis of present-tense fiction. *Language and Literature: International Journal of Stylistics* 28(3): 280–304.

Jucker AH (2015a) Pragmatics of fiction: literary uses of *uh* and *um*. *Journal of Pragmatics* 86: 63–67. DOI: 10.1016/j.pragma.2015.05.012.

Jucker AH (2015b) *Uh* and *um* as planners in the *Corpus of Historical American English*. In: Taavitsainen I, Kytö M, Claridge C, et al. (eds) *Developments in English: Expanding Electronic Evidence*. Cambridge: Cambridge University Press, pp. 162–177.

Lambrou M (2014) Stylistics, conversation analysis and the cooperative principle. In: Burke Michael (ed) *The Routledge Handbook of Stylistics*. Oxon: Routledge, pp. 136–154.

Landert D (2021) The spontaneous co-creation of comedy: humour in improvised theatrical fiction. *Journal of Pragmatics* 173: 6–87. DOI: 10.1016/j.pragma.2020.12.007.

Leech G, Hundt M, Mair C, et al. (2009) Change in contemporary english. A grammatical study. (Studies in English Language). Cambridge: Cambridge University Press.

Locher MA and Jucker AH (eds) (2017) *Pragmatics of Fiction*. (Handbooks of Pragmatics 12). Berlin/New York: De Gruyter Mouton.

Locher MA and Jucker AH (2021) *The Pragmatics of Fiction. Literature, Stage and Screen Discourse*. Edinburgh: Edinburgh University Press.

Norrick NR (2011) Interjections. In: Andersen G and Aijmer K (eds) *Pragmatics of Society.* (Handbooks of Pragmatics 5). Berlin: de Gruyter, pp. 243–292.

Norrick NR (2015) Interjections. In: Aijmer K and Rühlemann C (eds) *Corpus Pragmatics*. A Handbook. Cambridge: Cambridge University Press, pp. 249–275.

Oesterreicher W (1997) Types of orality in text. In: Bakker E and Kahane A (eds) *Written Voices, Spoken Signs. Tradition, Performance, and the Epic Text*. Cambridge, MA/London: Harvard University Press, pp. 190–214.

Quaglio P (2009) *Television Dialogue. The Sitcom* Friends *vs. Natural Conversation*. (Studies in Corpus Linguistics 36). Amsterdam/Philadelphia: John Benjamins.

Schiffrin D (1987) *Discourse Markers*. Studies in Interactional Sociolinguistics, 5. Cambridge: Cambridge University Press.

Short M (1996) *Exploring the Language of Poems, Plays and Prose*. London: Longman.

Stange U (2016) *Emotive Interjections in British English. A Corpus-Based Study on Variation in Acquisition, Function and Usage*. (Studies in Corpus Linguistics 75). Amsterdam: John Benjamins.

Staley L and Jucker AH (2021) "The uh deconstructed pumpkin pie": the use of *uh* and *um* in Los Angeles restaurant server talk. *Journal of Pragmatics* 172: 21–34. DOI: 10.1016/j.pragma.2020.11.004.

Thomas BE (1997) 'It's good to talk'? An analysis of a telephone conversation from Evelyn Waugh's *Vile Bodies*. *Language and Literature: International Journal of Stylistics* 6(2): 105–119.

Thomas BE (2002) Multiparty talk in the novel: the distribution of tea and talk in a scene from Evelyn Waugh's *Black Mischief*. *Poetics Today* 23: 657–684.

Tottie G (2011) *Uh* and *Um* as sociolinguistic markers in British English. *International Journal of Corpus Linguistics* 16: 173–197.

Tottie G (2014) On the use of *uh* and *um* in American English. *Discourse linguistics: Theory and practice* 21(1): 6–29.